Egan, Gabriel. 2007k. "'Electronic Publishing: Politics and Pragmatics': A paper delivered at 'Models of Partnership in Digital Research', a one-day colloquium at Sheffield Hallam University on 3 September."

"Electronic Publishing: Politics and Pragmatics" by Gabriel Egan

In 1958, America and Canada formed the North American Air Defense Command (NORAD) to coordinate data from radar stations monitoring the airspace over the North Pole. Contrary to Vannevar Bush's prediction, InterContinental Ballistic Missiles had turned out to be entirely feasible and in response to this new threat a NORAD command and control centre was built inside the hollowed-out Cheyenne Mountain next to Peterson Air Force Base in Colorado. The new weapons, ICBMs, were too fast to be tracked by the existing computers and through the 1960s the military's need for faster processing and secure communications across the continent drove research in electronics. The breakthrough came from a reproduction technology invented for commercial book publishing and refined by into a fine-art by Henri de Toulouse-Lautrec and Paul Gauguin: lithography. Rather than wiring individual components together, transistors, diodes, capacitors, and resistors of microscopic size could be lithographically printed directly onto semiconductor material, to make an Integrated Circuit, or micro-chip. In a delightful irony, the technology of printing remains at the heart of the digital revolution.

A second breakthrough was in networking. In the early 1960s Paul Baran of the Rand Corporation, a non-profit research and development organization funded by the American government (mostly the military), invented--or, more precisely, made explicit the application of--two techniques that would allow a military communications network to continue working efficiently even after many of its nodes and its links had been destroyed: packet switching and distributed networking. Baran's research appeared as a collection of Rand Memoranda reports for the American Air Force, deposited at the Defense Documentation Center in Virginia, and was subsequently published by the Institute of Electrical and Electronic Engineers (Baran 1964).

The packet switching method broke all communications into small, regularly formatted, units, each of which was sent off to find its destination independently and by whatever route it could. The distributed network model for connecting computers eschewed the common-sense (and militarily familiar) principle that each node (computer centre) had to be highly reliable and had to have a highly reliable link to a few others, and instead proposed that many relatively unreliable nodes should have many relatively unreliable links to many others. In the event of nodes and links failing, the intelligent and on-the-fly routing around the sites of failure would ensure that the network as a whole continued to work even when substantial proportions of it had been destroyed. The Advanced Research Projects Agency Network (ARPANET) embodying this technology first connected 4 university computer sites in 1969, and in 1972 it explicitly acquired the adjective 'defense' to become DARPANET, which became the Internet. Whilst it is not quite true that the Internet was invented to provide military communications that could survive a surprise nuclear attack--and hence enable an American retaliatory nuclear strike--the collaboration between military and academic research that Vannevar Bush initiated made such convergence virtually inevitable. Thus seemingly rational people were openly planning for reprisal in the event that hundreds of millions of their fellows were murdered with the new weapons technology.

Yet, in the late 1960s universities across the western world were places of radical challenge to prevailing ideas, including orthodoxies about what universities themselves were for. The Network Working Group that steered the academic side of ARPANET began to record its internal memoranda as a series of numbered documents, each of which was a Request for Comment (RFC) rather than an assertion of any particular power relation or the exercize of any particular rhetorical strategy. The third of these was dated April 1969 and self-reflexively concerned document conventions, encouraging informality in order "to promote the exchange and discussion of considerably less than authoritative ideas", and it noted its own self-contradictoriness: "These standards (or lack of them) are stated explicitly . . ." (RFC-3). The informality of graduate students was here in the service of academic research that was inherently also military research, showing the apparent natural affinity of these activities that Bush had ascertained before the end of the second world war. The protocols of the Internet are still proposed and agreed by the Request for Comments process, now numbering nearly 4,000.

*

We are here to talk about models of digital partnership, and my instinctive response as a contrary literary critic is of course to question whether partnership is the something that we should be concerned with, let alone to try to model. At least, I wonder whether we should think so much about the partners: the Joint Information Systems Committee (JISC), the big digital publishers, the international consortia in our research areas. If we think about what developments have been most visibly successful in the digital media recently, it is not the partnerships of big organizations at the top of the digital hierarchies, but the work down at the bottom of the pyramid. So, let me float for a minute a different approach to the digital, one not put together by big institutions and imposed from on high, but one arises from the grassroots. To illustrate it, I want to return to an acecdote from the second world war. Asked to design a new university campus in the 1930s, the German architect (later the wartime armaments minister) Albert Speer ordered that the green spaces between buildings should not be covered with paved walkways until one year after the building opened. 'Let us see where people actually walk, and then pave the paths they have chosen', Speer explained. The dominant tagging model for the production of digital versions of literary texts are the guidelines of the Text Encoding Initiative (TEI), a project initiated by the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing, and now spun off as an independent consortium that exists to maintain the TEI tagging standard. What a perfect model of digital partnership. TEI provides a standard set of tags by which anyone who wishes to make a digital version of a written work may mark-up the text in order to record the features that the later users of the text might wish to know about. Whereas a text printed on paper will use typographical and book-design conventions (say, changing to an sloping typeface for proper nouns and starting a fresh page for each new chapter), an intelligently tagged electronic text will explicitly mark structural and semantic features of the writing.

TEI aims to offer conventions whereby the maker of a digital version of text can record any feature that might conceivably be of interest to the subsequent users of the text. This is essentially an effort of preemption: the creator attempts to think ahead of the user and to tag all the features she could want to work with. This is not the only way to proceed, and I'd like to ponder the alternative way of tagging large digital corpora called *folksonomy*. Rather than produce a classificatory system, a taxonomy, in advance of the

users' access to a text, a *folksonomy* embodies Speer's insight about paths: it were better to wait until you have seen what users actually do with texts, and let their choices drive the process. Users of the classificatory systems embodied in the website systems *deli.cio.us* and *Flickr* are invited to describe things--websites, online articles, uploaded digital photographs--using tags created by other users rather than tags imposed in advanced by the inventors of the system. (Thence the *folks-* part of the word *folksonomy*.) In such systems, the tags that are first offered to the user are the ones that preceding users have most often chosen or have most often searched for, and this represents a practical expression of the principle that how people actually categorize things, how they actually carve-up reality in their classificatory attempts to make sense of it, ought to be the main determinant of how tagging is done. Might we not, I wonder, explore the possibility that we might employ such a *folksonomic* approach to the tagging of literary texts in the making of digital archives. There are all sorts of problems with *folksonomy* tagging, such as the control of synonyms, ambiguities (how to distinguish the tag the town 'Reading' from the tag for the activity of 'Reading', which is spelt the same) the adifficulty of generating hierarchies of tagging so that small things (varieties of tea, for example) are contained with larger things (classes of beverage, say). But perhaps the rapid expansion of *folksonomies* tells us that these disadvantages matter less than the exponential growth of metadata that follow from letting the readers decide for themselves.

Just as the world is littered with failed digital projects, it is littered with taxonomies that died. When the card catalogues of university libraries were being digitized in the 1970s, 80s, and 90s, it was widely objected that the paper versions embodied knowledge that would be lost of in the database versions. The library that I use most often is concerned almost entirely with the work of one author, Shakespeare, and so necessarily its books occupy one small part of the available classificatory namespace. Indeed, the Shakespeare Institute librarians had found they simply could not stick to the standard Library of Congress classification system because, as finely reticulated as it is, it nonetheless gave the same classmark to many hundreds of books. The card catalogue of the Shakespeare Institute library contained all kinds of clever fixes to adapt Library of Congress classification to their own peculiar circumstance and it was feared that the loss of this card catalogue would entail loss of vital knowledge about the collection. As it has turned out, the advantages of digitization have greatly outweighed these losses, and this is a valid general principles for taxonomies: even quite badly executed digitizations are better than carefully thought-out attempts to preserve knowledge in pre-digital forms, and indeed I suspect that this may be true of digital taxonomies too. I would argue, incidentally, that the same is true of book indexes: these simply become redundant once you have the full-text of a book. In a recent exchange of letters on this topic in the *Times Literary Supplement*, professional indexers insisted that the carefully put together human classifications that a hand-made index embodies would always be more useful to readers than computer-made concordances. My response is that those saying this simply haven't seen how clever the latest natural language processing software is in its morphosyntactic analysis of written English. Indeed, I think those who are sure that readers will always prefer well-made manual indexes are like the medieval scribes who, confronted by the crude but rapid early printing press, comforted themselves that serious readers would always prefer painstakingly illuminated manuscripts.

If I were to generalize the points I've made so far, then, I would say that quick-and-dirty, rather than slow and methodical, is generally the way to go in digital projects, that top-

down classifications are less useful than user-led, even anarchic, categorizations, and that there is a kind of analogy here in the very technical basis of the Internet. That is to say, just as multiple, redundant, unreliable connections are better than a few strong and efficient ones, multiple, disorganized taxonomies may well be better than a few highly structured ones, like TEI.

*

Which brings me to a larger point, which is the digital publishing has not developed in the ways that were anticipated. The worldwide-web was invented to be a means of writing as much as reading: Tim Berners-Lee's idea was that those who want to communicate would author their stuff in HTML and self-publish it on the Internet by linking in to other writings. It hasn't worked out that way, at least not until recently. Almost all writers use the closed, proprietary format of Microsoft Word rather than the open, free standard HTML. Moreover, even when making online presences most individuals don't use HTML, at least not directly. The very recent big explosion of self-publishing by individuals has been using visual editors that run on remote servers, hiding the raw HTML from the user and enabling people to build sites directly on the server that will host them rather than doing the two-stage process of writing their material locally and then sending it to the server to make it visible on the web. I'm referring here to the extremely popular social networking sites (Facebook, MySpace) that enable people to have an online presence without using HTML editors. This process is extending into higher education in the UK, and I suspect in other countries too, as Virtual Learning Environment packages take over from open-standard educational websites. Indeed, even institutions that have avoided proprietary software are going this way because the the open-source Virtual Learning Environment called Moodle convinces them to give up on using HTML directly. At my institution, the HTTP webspace is strictly controlled and can only be edited by authorized technicians. The Moodle networl, though, is open to all staff and easy to edit with, so that is increasingly where we put all materials for students such as timetables and forms for choosing modules.

The technology, then, is only now starting to make readers also writers--it's taken much longer than we expected. What about academic writers? The technologies, economics, and politics of scholarly publication in the humanities look set to change rapidly in the near future. Even if the market for print publication were to remain relatively buoyant, national governments (the main direct and indirect funders of research) are increasingly questioning the efficiency and cost of traditional means of dissemination. The academic humanities book market is an unusual sector in publishing because the producers, the academic authors, comprise also the largest sector of the consumers, either directly or through their institutional libraries. From the perspective of those who pay for research, publishers appear to have created and plugged themselves into a circuit of knowledge dissemination (from academics to publishers and back again) to which they do not contribute as much value as they extract. With new electronic publication technologies that do not require large investments of capital (printing presses, warehouses, transport), there are powerful forces directing academic authors away from traditional print publication. At Ray Siemen's suggesting, I've begun putting together a collection of essays that will bring together a team of academics with experience in this new field in order to explore not only the practical matters of electronic publication but also to have them reflect on the politics of the vastly changed knowledge landscape that is likely to exist soon. As well as eliciting their accounts of how such matters as Intellectual

Property Rights (IPR) and coding standards figured in their projects, the essays will draw from the contributors their wider visions of the future of the knowledge economy and how the humanities disciplines will fare in a world that increasingly trusts its cultural heritage to magnetism and laser optics rather than inks and paper.

The essays will be grouped into two categories, concerning the creation of digital projects and their dissemination. There is always a necessary connection between writing and reading--acts of production presume their own subsequent consumption--but in this topic the linkage is especially close because the technology is now at last starting to narrow the gap between 'making' and 'getting'. That is to say, in print publication the author uses creative tools (paper, typewriters, computers) that are nothing like the tools for dissemination (presses and binding machines, trucks/boats/planes, warehouses, and retail outlets). In electronic publishing, however, the author creates very nearly the final object and the act of distribution alters it little or not at all. When materials are created directly on the servers that make them available to the worldwide web rather than being sent by File Transfer Protocol, the circle has been finally closed. Whereas in print publication the author may leave the final choices of mediation (paper quality, binding, pagination, image reproduction) to someone else, in electronic publishing the creator is obliged to consider closely how the reader will experience the content. For this reason, contributors who have made digital content will have interesting things to say about how their close involvement in the dissemination shaped their creations. This narrowing of the gap between producer and consumer is also crucial in regard to the adoption of standards since creators must ensure their output will work with the computers available to consumers: there is no point planning a project whose outputs are a LaserDisc with accompanying 8-track audio-book.

The recent past of electronic publication is littered with projects that have become unusable because of rapid changes in the standards of computer hardware and software. The BBC's 1980s project to create a new digital Domesday Book recorded life in the United Kingdom 900 years after the original book. The many thousands of items assembled for this project are now lost because the hardware platform, a LaserDisc attached to an Acorn/BBC micro-computer, is incompatible with standard computers today. The 950-year old first (paper) edition of the book remains readable and the digital version is useless. The widespread adoption of computer standards ratified by the International Standards Organization (ISO) ameliorates but does not eliminate this problem.

In the digitization of writing the Text Encoding Initiative (TEI) aims to provide a standard set of tags, and in all disciplines Extensible Markup Language is touted as the means to ensure that intellectual content can be migrated across the ever-changing future hardware and software standards of the computer marketplace. Nonetheless, debates continue about the philosophical underpinnings of such standards and, in the case of TEI, whether creators should tag only the logical structures of writing (line, stanza, chapter, and so on) and not its visual appearance (boldface, italic, whitespace). Despite the widespread acceptance of TEI the tools needed to produce an electronic edition are considerably more complex and fragile than those needed to produce a print edition (for which Microsoft Word is ubiquitous) and within certain learned societies the discipline of tagging is considered solely the young scholars' province because it is so technically difficult. One school of thought maintains that careful human tagging of old texts will turn out to be wasted effort because computers will eventually be able to make sense of texts

themselves. The essays in this section of the book will survey the creation of electronic publications from the contributors' perspectives, attempting to point the way forward. The contributors will be asked to go beyond merely recounting their experiences in order to think about the fundamental differences in the production/consumption dialectic that are entaild in 'born digital' content generation, and to suggest possible future relations of creators (whom we used to call authors) and consumers (whom we used to call readers) in humanities research. Just as there are projects that failed because computer standards moved on, there are projects that failed because too few buyers were willing to pay for the product on sale. At several thousands of dollars apiece, almost no-one bought the Arden Shakespeare CD-ROM or the Thomson Gale English Short Title Catalog (ESTC) on CD-ROM. The print version of the Encyclopedia Britannica costs about that much and has remained just about marketable, while the CD-ROM version of the same content has settled at around $50. Unusually in this sector, the Oxford English Dictionary appeals to a large non-specialist market and has been able to keep its price relatively high at around $500. By contrast, the market for the specialist research book is small and the price-per-unit is fairly high (around $100-200). This market may not be sustainable if academics and research-library buyers cease to support it, as seems distinctly possible.

Since the Budapest Initiative in 2002, the Bethesda Statement in 2003, and the Berlin Declaration in 2003, the idea that the results of scholarly research should be given away freely over the Internet--the Open Access (OA) principle--has gained many adherents. The United Kingdom government's Science and Technology Committee has declared itself in favour of OA and endorsed three routes: 1) the creation of Institutional Repositories (IRs) to hold and preserve research outputs from particular institutions (primarily universities); 2) self-archiving (web-based dissemination) by individual academics; 3) Author-pays publishing instead of reader-pays as at present. OA has been driven by the journal-centric sciences, but the humanities disciplines place at least as much importance on the book as the journal article. There may be book-specific barriers to OA and although major academic publishers (such as Elsevier and Taylor & Francis) have declared their acceptance of some OA principles in respect of journal articles (limited dissemination and deposit of pre-print articles) the picture with monographs and critical editions is far from clear.

One section of the book I'm editing will be concerned with reader-pays (traditional) electronic publishing as well as Open Access and I am inviting contributors to reflect on their experiences of these two modes and to consider the benefits and demerits of each. Topics covered may include the degree to which academic institutions ought to supplant publishers by offering primary research outputs in their Institutional Repositories, and how far (if at all) they should be able to compel academics to deposit their works in the IR. Might whole books become available in IRs? Those with experience of hybrid print-and-electronic publications will be invited to describe how the electronic version was conceived and at what stage in the commissioning it came up for discussion. Was there any suggestion that the electronic version might be OA instead of a priced product? Is the electronic version a supplement to the print version, or is it a dumping ground for material that cannot be crammed into the book? Because the Worldwide Web contains many unreliable sources in humanities disciplines, the electronic medium has suffered a credibility gap in comparison with print resources. Contributors will be invited to discuss whether they find themselves mistrusting electronic media and whether questions of permanence (the text and pictures never changing) are a barrier to the media's

acceptance. In relation to professional advancement based on peer review, is it possible that the bypassing of publishers will lead to better quality controls as seems to have happened in the sciences? Those with an interest in the economics of publishing might like to consider the viability of the author-pays model of publishing in the humanities, and the economic consequences of fundamental changes in property law and rights that seem entailed by the new media. The humanities differ from the sciences in their valuing of old materials--a book written 100 years ago may still be the last word on a topic--and so the problems of media longevity and knowledge archiving are considerable more acute for us. Contributors may wish to consider the argument that for many purposes we should remain wedded to paper for the forseeable future.

Where does that leave us as academics, especially those of us who do their primary research in libraries and using books, and who produce research output also in the form of books. It is a peculiar situation, of course. We are the suppliers of monographs and the consumers. This is not true of trade books, reference books, and text-books, and I leave those aside from these comments. In the decade 1995-2004 the share of UK's Higher Education library budget spent on monographs dropped from 45% to 35% and yet the number of monographs published doubled. Each title sells fewer copies and libraries--especially those hit by the spiralling costs of serials--cannot afford to buy the same portion of all that gets published. Monograph print runs in 1960s were 5 to 10 times those today, and the big publishers such Macmillan made a lot of money. Some Oxford University Press monographs are produced in runs of just 200, so effectively only the preorders are printed for. In these straitened time we see the consolidation of publishers: Wiley has taken over Blackwells and no longer does monographs, and Routledge has become part of Taylor and Francis. There is no diminution in the writing of books, and in the UK the Research Assessment Exercise--a one-off audit of research output--has produced a massive glut of books that academics had to write for career advancement but that very few people want to buy.

In these competitive markets, as print runs shorten there have had to be cuts in production costs: less careful copy editing, less careful manual typesetting. One hope for the publishers seems to be print-on-demand, the Just in Time solution to their warehousing costs. This has made works previously impossible now possible, but so far only backlists have been put 'on demand'. Why not front lists? With print-runs as short as Oxford University Press's 200 copies for the preorder market, this is in effect print-on-demand. Ray was talking about Wal-Mart getting rid of warehousing and thereby being able to beat the competition. In the case of publishers, it's not clear that even with the advantage of removing the cost of warehousing, any of them can survive. Once it becomes clear that publishers are really only possessors of electronic repositories of texts and do not have large amounts of capital tied up in print versions for the speculative market, it will increasingly seem absurd for academics and university libraries to help protect their business. More and more peoply will start to ask just what, in our electronic world, do the publishers think they are bringing to the party? That is one of the questions that I want contributors to this book to think about, and the answer might be that there is no place at all for the publishers.

*

How should we as academics react to all this? One thing I think we could usefully do is ignore copyright restrictions and copy any and all texts that we or our students might

want to use. It is no exaggeration to say that the new media are fundamentally altering the nature of property within late industrial capitalism, and that old notions of ownership simply do not apply in the new situations. The technology of almost instantaneous and absolutely perfect digital reproduction makes a mockery of laws written in the days when copying was painfully slow and never perfect. Indeed, the very impermanence of online resources puts us under a moral obligation to pirate as much as possible, because we cannot rely on the materials surviving any other way. As is well known, the BBC routinely wiped and reused tapes of radio and television programmes from the 1950s and 1960s, and in many cases the only surviving copies are illegal pirated recordings made off-the-air by listeners and viewers and stored at home. The BBC is now grateful to receive copies of these illegal recordings to fill the extensive gaps in its broadcasting archive. On a personal level, I'm sure I'm not the only person here whose list of publications includes an article commissioned for an academic website that no longer exists. In my case, the I only hope that (contrary to the terms of use published on the site) people did copy material from the Arden Shakespeare's now defunct ArdenNet website, else I'm the sole possessor of an text that was once widely available and that has been cited in more than one printed book.

In a world in which Google is routinely scanning books without their authors' permission--anybody who has not yet looked might be surprised to find out how much of their own stuff Google already has in digital form--and in which universities are seeking to put publishers out of business and make themselves repositories of knowledge in electronic form and in which large public institutions have shown themselves to be unreliable custodians of data, it would be an absurdly self-denying gesture for academics, the source of all this knowledge, to pause before copying materials and ponder the copyright position of their acts. It greatly surprises me how many people object when I make this argument, saying that I am promoting theft. The key defining attribute of a theft is that it deprives the rightful owner of property her use of it. To copy a CD or DVD or downloaded digital file brings into existence a new object and leaves the original unchanged, so unlike the theft of a object the 'victim' is in no worse a position than she was before the crime was committed. This is not theft. Ordinary property has a tangible existence in the world and cultures across the world have for millennia enforced rules about its ownership. Intellectual Property is a relatively recent invention, is entirely intangible, and emerges from the particular configuration of the technologies of reproduction at a particular moment in history. In truth, Intellectual Property is not actually property at all, precisely because it cannot be stolen.